SWHID specification

kickoff meeting

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

March 27th 2023

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

What and how

Objective

bring you all up to speed on the SWHID specification effort, and kickstart the work

Organization

- duration: 1 hour (45m plus 15m Q&A); write questions in the chat
- session will be recorded
- set your Zoom id to your full name

Agenda

- recall basic notions of identifiers, then focus on the SWHID
- review of the way SWHIDs are composed and computed
- focus on the key parts of the specification that need work
- guided tour of the contribution and editorial process
- governance and licensing

Outline

Identification vs Location — an example is worth a thousand words

Identification of a book

Goal: identify a book

- one ISBN number per published book
- ISO 2108 standard specification

Location of (a copy of) a book

Goal: find (a copy of) a book

- many locations (locations can change!)
- many approaches for call numbers

we are interested in identification, not in location

Extrinsic vs Intrinsic identifiers

In a nutshell (for more info see this dedicated blog post)

Main difference: how the *relation* between *identifier* and *designated object* is created and maintained. *Persistence* is a key desired property.

	Extrinsic	**Intrinsic**
relation	register	convention
persistence	external[a]	internal
pre-internet	passport number, ISBN, SSN, etc.	Music/Chemistry notations *e.g. NaCl is table salt*
internet era	DOI, Handle, Ark, etc.	cryptographic hashes *e.g.: git, bitcoin, SWHID*

[a] "persistence... is a function of *administrative* care" RFC 3650 (Handle System Overview, 2003)

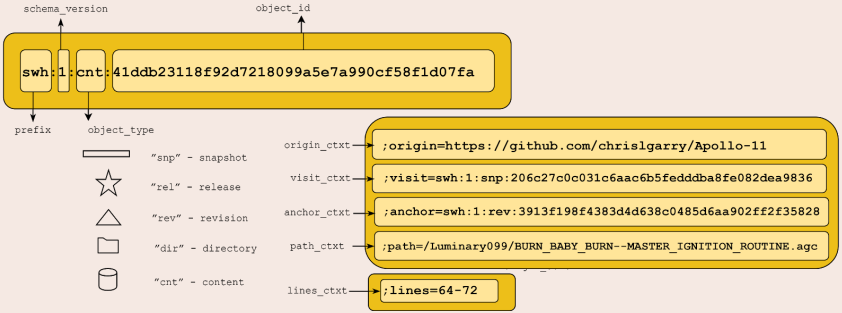
Here we are interested in normalising the SWHID *intrinsic identifier*

Outline

Bird's eye view of the SWHID *Intrinsic* Identifier

Structure of a SWHID identifier

link to full docs

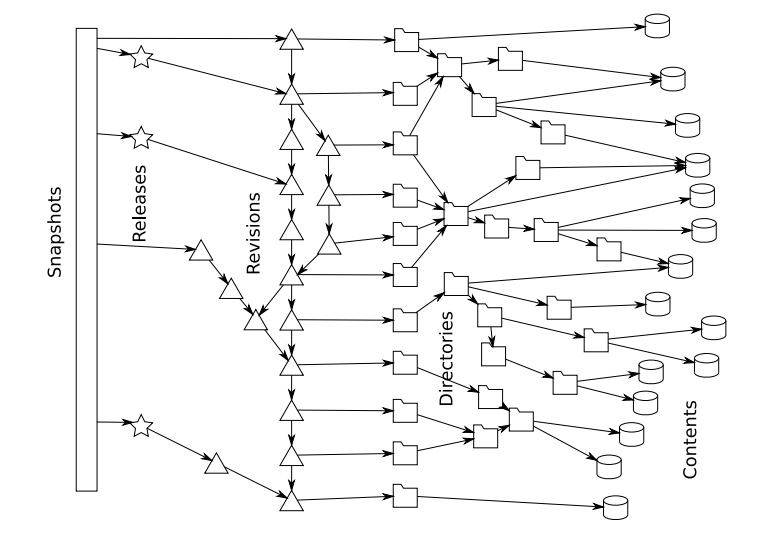


Current status

30+B SWHIDs in the Software Heritage archive

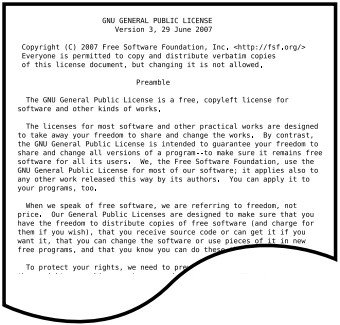
Mention in Linux Foundation's SPDX 2.2; IANA registered; WikiData P6138

SWHID computation: a worked example



SWHID computation: a worked example

Contents



GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

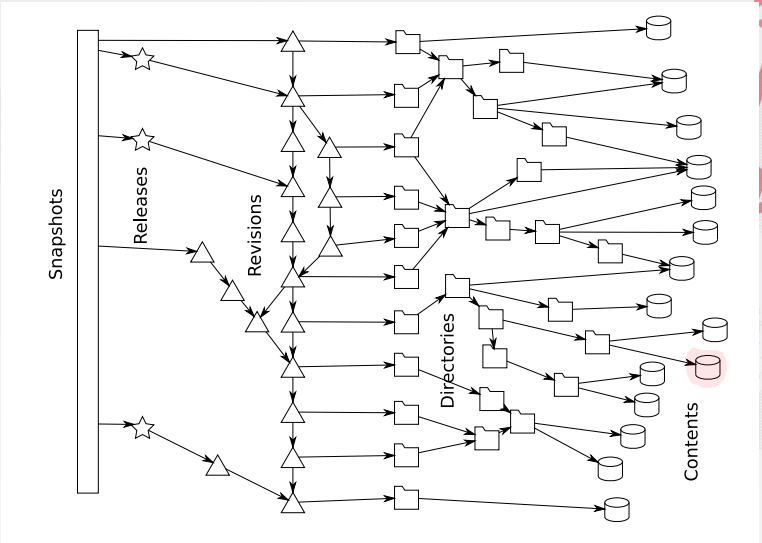
The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program--to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these

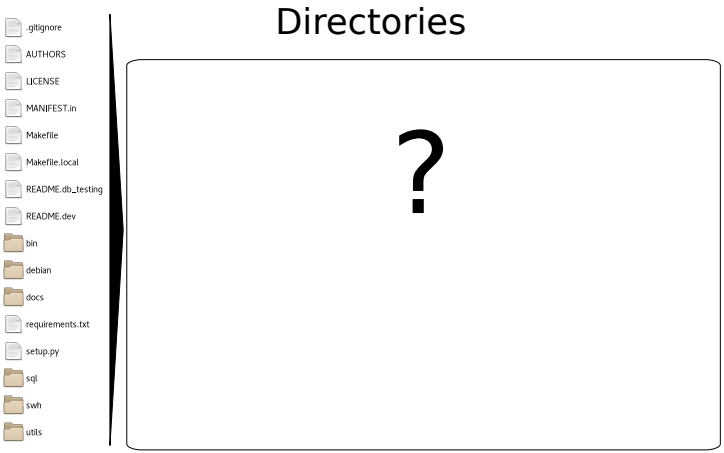
To protect your rights, we need to pre

sha1: 8624bcdae55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147

SWHID computation: a worked example



SWHID computation: a worked example



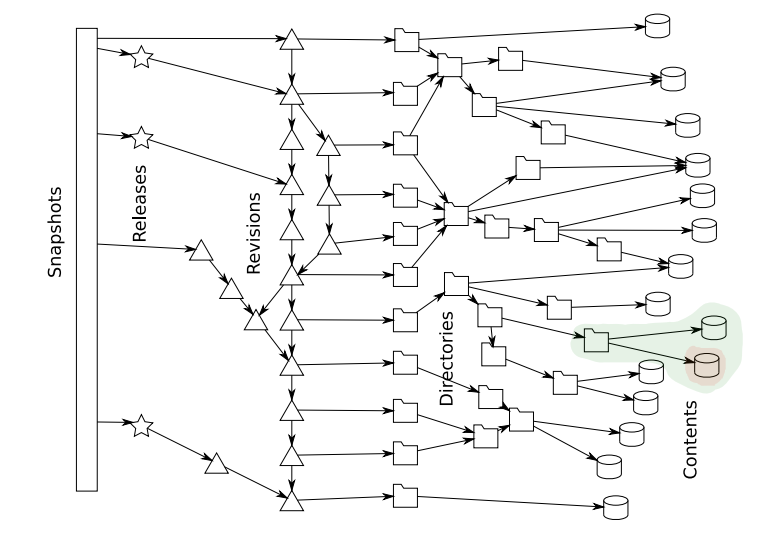
SWHID computation: a worked example

Directories

```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbddb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b085b3e07b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecef948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

SWHID computation: a worked example



SWHID computation: a worked example

Revisions

Details Changes Files

SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6

Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)

Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)

Subject: **provenance.tasks: add the revision -> origin cache task**

Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : *test_storage: properly pipeline origin and cont...*

provenance.tasks: add the revision -> origin cache task

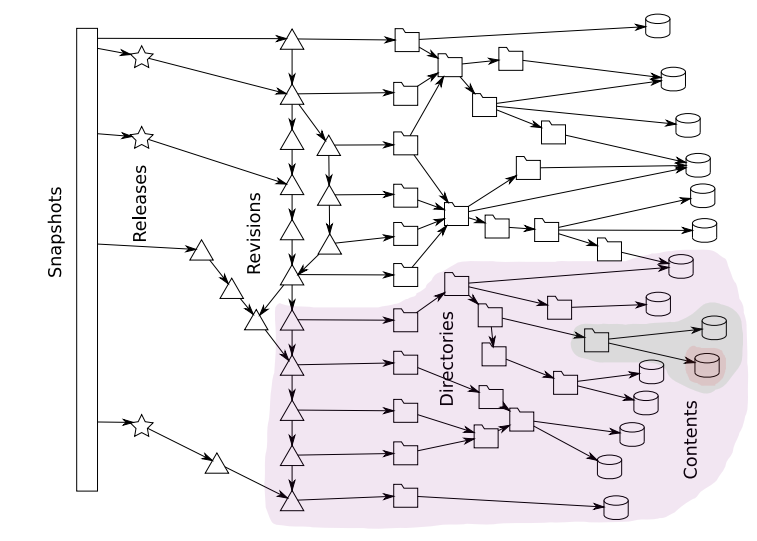
swh/storage/provenance/tasks.py 77

tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

provenance.tasks: add the revision -> origin cache task

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

SWHID computation: a worked example



SWHID computation: a worked example

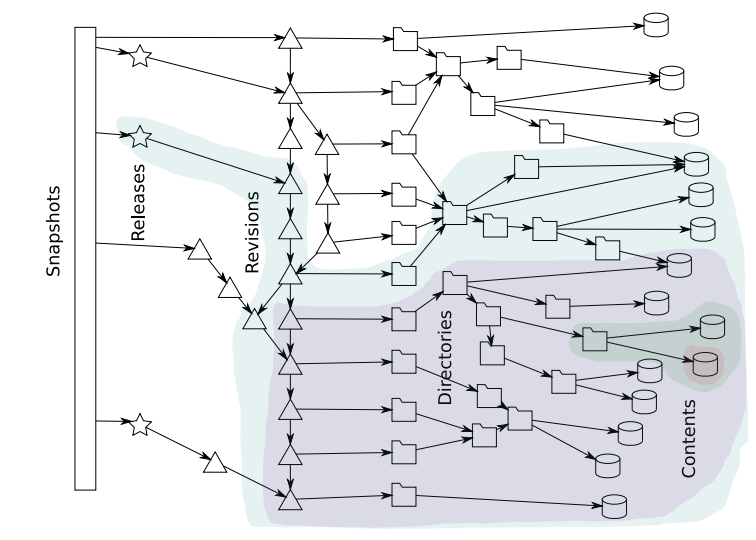
Releases

```
tag v0.0.51
Tagger: Nicolas Dandrimont <nicolas@dandrimont.eu>
Date:   Wed Aug 24 14:36:03 2016 +0200

Release swh.storage v0.0.51

 - Add new metadata column to origin_visit
 - Update swh-add-directory script for updated API
[...]

commit c0c9f16b1e134f593e7567570a1761b156e6eb1d
```

```
object c0c9f16b1e134f593e7567570a1761b156e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

Release swh.storage v0.0.51

 - Add new metadata column to origin_visit
 - Update swh-add-directory script for updated API
-----BEGIN PGP SIGNATURE-----

iQIzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqorw/aq6SOb5DijzEa+kWN3rXgVS+1K1vEVh1wNKAwx8eKJ7aX2kEiLDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
ICBIit2uJtXuCrDt93eKKPwvzZXg+hB0sMWy35Dr6jW7Z7K4Mu/PGglyIHPY55yo
IGEndWno7VfH1Vm6t1n5qB7I5mxRaqA+becqddubTZ2xjj+jpIUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBInPoS5TH0tujoJEVgPK/dHSP79QuHDHZFkCao
kIj6kAWyUB0Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1Il/g0EDfnsW67G6sDwKPKPHhgfVLQ3nV3GaQQTnu1RpMz006H9/tAwzC
Gg/K1PdHT4hzOiI46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrIJSUOMn
RpTTfUsbXUeXHGOpkgXhSYTnvp1gdPc76USTkK0aGe84AZm1Ik0mGrwXCVfPqlYo
nhhibBSHBNMoqyF6yTSOpUbYK7OtpYRRUGKWDeRKOwKSxkWKUZGtKzy6JYaJjo29
gulwqZQIf5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHIuUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

SWHID computation: a worked example



SWHID computation: a worked example

git show-refs

Snapshots

```
commit 08ffeb2577010925eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbe1c05d27238d9c5 refs/heads/foo
commit cf7ff9eea0eb22f8946908f5a8019f67de468e08 refs/heads/master
commit 7eca197fc66d2024047e54b1ed9e8b44361a0fc2 refs/heads/tmp-directory-add
commit 642a205f37de85005a85d427b53ee4fb2252e82e refs/heads/tmp/generic-releases
tag 20f043b1379cf768d9665977997fd4907c757f755 refs/tags/v0.0.1
tag 72a21991a384e539996dbb867bfb0bee72aee2cd refs/tags/v0.0.10
tag 3590e0ca0ebb070e5b376705fa230bbfa4ffa5cc refs/tags/v0.0.11
tag 33378427a403ba569a67777b8d58f6674fbc6556 refs/tags/v0.0.12
tag 06f74652755b327cf590311c2bfa036cf3b4b35d refs/tags/v0.0.13
tag 5a63255fe86ab854b581d744266f7d92a11e32f3bd refs/tags/v0.0.14
tag 586fba4e580b4f5fab05f599367643cbcb1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4098bf363177742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbed35efb202fee035c809abc7d6 refs/tags/v0.0.17
tag 228a2f1650dd12222e556559462e1e06fc4993d9 refs/tags/v0.0.18
tag 606979a4ca005d497fcd024aad00dce82636ef47c refs/tags/v0.0.19
tag 32bf5a59fc2a323baa6d5f15a6ad5382ec275a67 refs/tags/v0.0.2
tag 3147c3d31ec46cf6492f881e908b1237ebdff2c7 refs/tags/v0.0.20
tag 215ea50daba111e082e0b72e76eb4b6073a87908 refs/tags/v0.0.21
tag 3fb168c2072a5d6252124257a1e5dfc0f5ffa1df refs/tags/v0.0.22
tag 8cdbee8da4d73fc5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b

Outline

First step: the name of the game

Origin of the SWHID acronym

SWHIDs were born as

- intrinsic identifiers
- designed for
  - *software source code*
  - archived in Software Heritage

Hence the acronym:

SW Software

H Heritage

ID IDentifier

Proposed reformulation

SWHIDs are

- based on a *cryptographic hash*
- can be used *independently of the Software Heritage Archive*,
- *not restricted to source code*

Hence the proposal:

SW Software

H ~~Heritage~~ Hash

ID IDentifier

Setting the stage

Our goal

Create a *specification* that

- is *complete*, *precise* and *non ambiguous* (pictures are *simplified representations*!)
- allows any "person skilled in the art" to implement *the same calculation algorithm*

To this end we need to get right:

- five key parts in the core: *cnt, dir, rel, rev, snp*
- qualifiers: easier, but important too
- reference to external standards used (e.g. SHA algorithm)

What we have

- high level documentation from Software Heritage
- two complete implementations (one in Python for SWH, one in OCaml for Opam)
- a draft specification that needs to be completed

Governance and License

Bylaws

Detailed in the Governance Document

- consensus based decision making
- open process to produce an Approved Specification

License

specification Community Specification License

contributions specific CLA accepted by sign-off on GitHub

Coordination by the Core Team

Alexios Zavras, Jean-François Abramatic, Morane Gruenpeter, Roberto Di Cosmo, Stefano Zacchiroli

Roles: maintainers, editors

Documents, collaboration tools and workflow

Specification document

sources GitHub repository https://github.com/swhid/specification

rendered swhid.org website

Working on the specification

- contribution workflow based on *issues* and *pull requests* on GitHub
  - N.B: pull requests *must be signed-off*
- editors *merge* pull requests *on consensus*
- merges *rebuild* the rendered version on swhid.org

Mailing list: swhid-discuss

used mostly for coordination by the editors

(e.g. scheduling meeting, votes, version freeze, etc.)

Guided tour

Working with issues and pull requests

- the directory SWHID
  - issue
  - pull request

Taking a vote

- the name of the game poll is open at https://bit.ly/swhid-name-poll

Outline

Ready, set, go!

Timeline

- phase 1 complete and accurate v1.0 as soon as possible (end of April desirable)
  - help establish the whole process
  - cover *only* what is already known and being used
  - focus on items labeled blocker, make them *complete*, *precise* and *non ambiguous*
- phase 2 work towards v1.1
  - handle other feedback / input / requests
  - get the version candidate to become an ISO standard

Questions?

Links

- main enty point: https://swhid.org
- specification sources: https://github.com/swhid/specification